AP Statistics Chapter 4 – Designing Studies

4.1: Surveys and Samples

Population, Census and Sample

- The **population** in a statistical study is the entire group of individuals we want information about. For example, all registered voters in a given county.
- A **census** collects data from every individual in the population.
- A **sample** is a subset of individuals in the population from which we actually collect data.

Bias

The design of a statistical study shows **bias** if it would consistently underestimate or consistently overestimate the value you want to know.

Convenience Sampling

A **convenience sample** chooses the individuals easiest to reach. This will typically result in a biased sample of like-minded individuals.

Voluntary Response Sample

A **voluntary response** sample consists of people who choose themselves by responding to a general invitation. Voluntary response samples show bias because people with strong opinions (often in the same direction) are most likely to respond.

Simple Random Sample

A **simple random sample** (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

Random Digits

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these two properties:

1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9.
2. The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Choosing an SRS

Choose an **SRS** in two steps:

Step 1: Label. Assign a numerical label to every individual in the population.

Step 2: Table. Use Table B to select labels at random.

Stratified Random Sample

To get a **stratified random sample**, start by classifying the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the sample.

Cluster Sample

To get a **cluster sample**, start by classifying the population into groups of individuals that are located near each other, called *clusters*. Then choose an SRS of the **clusters**. All individuals in the chosen clusters are included in the sample.

Forms of Bias in Surveys and Samples

- **Undercoverage** occurs when some members of the population cannot be chosen in a sample.
- **Nonresponse** occurs when an individual chosen for the sample can't be contacted or refuses to participate.
- A systematic pattern of incorrect responses in a sample survey leads to **response bias**.
- The **wording of questions** is the most important influence on the answers given to a sample survey.

4.2: Experiments

Observational Study vs Experiment

- An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.
- An **experiment** deliberately imposes some treatment on individuals to measure their responses.
- When our goal is to understand cause and effect, experiments are the *only* source of fully convincing data. The distinction between observational study and experiment is one of the most important in statistics.

Confounding occurs when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other. Observational studies often fail to provide valid causal links between variables due to confounding

The Language of Experiments

A specific condition applied to the individuals in an experiment is called a **treatment**. If an experiment has several explanatory variables, a treatment is a combination of specific values of these variables.

The **experimental units** are the smallest collection of individuals to which treatments are applied. When the units are human beings, they often are called **subjects**.

Principles of Experimental Design

The basic principles for designing experiments are as follows:

1. **Comparison**. Use a design that compares two or more treatments.
2. **Random assignment**. Use chance to assign experimental units to treatments. Doing so helps create roughly equivalent groups of experimental units by balancing the effects of other variables among the treatment groups.
3. **Control**. Keep other variables that might affect the response the same for all groups.
4. **Replication**. Use enough experimental units in each group so that any differences in the effects of the treatments can be distinguished from chance differences between the groups.

Statistical Significance

- An observed effect so large that it would rarely occur by chance is called **statistically significant**.
- A statistically significant association in data from a well-designed experiment *does* imply causation.

Completely Randomized Design

- In a **completely randomized design**, the treatments are assigned to all the experimental units completely by chance.
- Some experiments may include a **control group** that receives an inactive treatment or an existing baseline treatment.
- The response to a dummy treatment is called the **placebo effect**.
- In a **double-blind experiment**, neither the subjects nor those who interact with them and measure the response variable know which treatment a subject received.

Block Design

A **block** is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.
In a **randomized block design**, the random assignment of experimental units to treatments is carried out separately within each block.

Matched Pairs Design

- A **matched pairs design** is a randomized blocked experiment in which each block consists of a matching pair of similar experimental units.
- Chance is used to determine which unit in each pair gets each treatment.
- Sometimes, a "pair" in a matched-pairs design consists of a single unit that receives both treatments. Since the order of the treatments can influence the response, chance is used to determine with treatment is applied first for each unit.

AP Statistics Chapter 4 Study Guide

Multiple Choice
Identify the choice that best completes the statement or answers the question.

1. Can pleasant aromas help a student learn better? Two researchers believed that the presence of a floral scent could improve a person's learning ability in certain situations. They had 22 people work through a pencil-and-paper maze six times, three times while wearing a floral-scented mask and three times wearing an unscented mask. The three trials for each mask closely followed one another. Testers measured the length of time it took subjects to complete each of the six trials. They reported that, on average, subjects wearing the floral-scented mask completed the maze more quickly than those wearing the unscented mask, although the difference was not statistically significant. This study is
   a. a convenience sample.
   b. an observational study, not an experiment.
   c. an experiment, but not a double-blind experiment.
   d. a double-blind experiment.
   e. a voluntary response sample.

2. A television station is interested in predicting whether voters in its viewing area are in favor of federal funding for abortions. It asks its viewers to phone in and indicate whether they support/are in favor of or are opposed to this. Of the 2241 viewers who phoned in, 1574 (70.24%) were opposed to federal funding for abortions.

   Referring to the information above, the viewers who phoned in are
   a. a voluntary response sample.
   b. a convenience sample.
   c. a probability sample.
   d. a population.
   e. a census

3. A marketing research firm wishes to determine if the adult men in Laramie, Wyoming, would be interested in a new upscale men's clothing store. From a list of all residential addresses in Laramie, the firm selects a simple random sample of 100 and mails a brief questionnaire to each. The population of interest is
   a. all adult men in Laramie, Wyoming.
   b. all residential addresses in Laramie, Wyoming.
   c. the members of the marketing firm that actually conducted the survey.
   d. the 100 addresses to which the survey was mailed.
   e. upscale men's clothing stores

4. In order to assess the opinion of students at the University of Minnesota on campus snow removal, a reporter for the student newspaper interviews the first 12 students he meets who are willing to express their opinion. In this case, the sample is
   a. all those students favoring prompt snow removal.
   b. all students at universities receiving substantial snow.
   c. the 12 students interviewed.
   d. all students at the University of Minnesota.
   e. reporters at the student newspaper

5. A researcher is interested in the cholesterol levels of adults in the city in which she lives. A free cholesterol screening program is set up in the downtown area during the lunch hour. Individuals can walk in and have their cholesterol levels determined for free. One hundred and seventy three people use the service, and their average cholesterol is 217.8. The sample obtained is an example of
   a. a simple random sample, since the experimenter did not know beforehand which individuals would come to the screening.
   b. a stratified sample of high and low cholesterol individuals.
   c. a sample probably containing bias and undercoverage.
   d. a multistage sample of varying cholesterol levels.
   e. a systematic random sample

6. You are testing a new medication for relief of depression. You are going to give the new medication to subjects suffering from depression and see if their symptoms have lessened after a month. You have eight subjects available. Half of the subjects are to be given the new medication and the other half a placebo. The names of the eight subjects are given below.

1. Blumenthal	5. House
2. Costello	6. Long
3. Duvall	7. Pavlicova
4. Fan	8. Tang

Using the list of random digits

```
81507 27102 56027 55892 33063 41842
81868 71035 09001 43367 49497
```

starting at the beginning of this list and using single-digit labels, you assign the first four subjects selected to receive the new medication, while the remainder receive the placebo. The subjects assigned to the placebo are
   a. Blumenthal, Costello, Duvall, and Fan.
   b. Blumenthal, House, Pavlicova, and Tang.
   c. House, Long, Pavlicova, and Tang.
   d. Costello, Duvall, Fan, and Long.
   e. Costello, House, Duvall and Long

7. A study of human development showed two types of movies to groups of children. Crackers were available in a bowl, and the investigators compared the number of crackers eaten by children watching the different kinds of movies. One kind of movie was shown at 8 AM (right after the children had breakfast) and another at 11 AM (right before the children had lunch). It was found that during the movie shown at 11 AM, more crackers were eaten than during the movie shown at 8 AM. The investigators concluded that the different types of movies had an effect on appetite.

   The results cannot be trusted because
   a. the study was not double-blind. Neither the investigators nor the children should have been aware of which movie was being shown.
   b. the investigators were biased. They knew beforehand what they hoped the study would show.
   c. the investigators should have used several bowls, with crackers randomly placed in each.
   d. children do not eat crackers while watching movies
   e. the time the movie was shown is a confounding variable.

8. We say that the design of a study is biased if which of the following is true?
   a. A racial or sexual preference is suspected
   b. Random placebos have been used
   c. Certain outcomes are systematically favored
   d. The correlation is greater than 1 or less than –1
   e. None of the above.

9. Consider an experiment to investigate the effectiveness of different insecticides in controlling pests and their effects on subsequent yield. What is the best reason for randomly assigning treatment levels (spraying or not spraying) to the experimental units (farms)?
   a. Randomization makes the experiment easier to conduct since we can apply the insecticide in any pattern rather than in a systematic fashion.
   b. Randomization will tend to average out all other uncontrolled factors such as soil fertility so that they are not confounded with the treatment effects.
   c. Randomization makes the analysis easier since the data can be collected and entered into the computer in any order.
   d. Randomization is required by statistical consultants before they will help you analyze the experiment.
   e. Randomization implies that it is not necessary to be careful during the experiment, during data collection, and during data analysis.

10. A nutritionist wants to study the effect of storage time (6, 12, and 18 months) on the amount of vitamin C present in freeze dried fruit when stored for these lengths of time. Vitamin C is measured in milligrams per 100 milligrams of fruit. Six fruit packs were randomly assigned to each of the three storage times. The treatment, experimental unit, and response are respectively:
    a. A specific storage time, amount of vitamin C, a fruit pack
    b. A fruit pack, amount of vitamin C, a specific storage time
    c. Random assignment, a fruit pack, amount of vitamin C
    d. A specific storage time, a fruit pack, amount of vitamin C
    e. A specific storage time, the nutritionist, amount of vitamin C